AN UPPER BOUND FOR THE HALES–JEWETT NUMBER HJ(4, 2)

MIKHAIL LAVROV

Abstract. We show that for n at least 10^{11}, any 2-coloring of the n-dimensional grid $[4]^n$ contains a monochromatic combinatorial line. This is a special case of the Hales–Jewett Theorem [4], to which the best known general upper bound is due to Shelah [6]; Shelah's recursion gives an upper bound between $2 \uparrow\uparrow 7$ and $2 \uparrow\uparrow 8$ for the case we consider, and no better value was previously known.

1. Introduction

Consider r-colorings of the n-dimensional grid $[t]^n = \{1, 2, \ldots, t\}^n$. We define a *combinatorial line* in $[t]^n$ to be an injective function $\ell : [t] \to [t]^n$ such that for each coordinate $1 \leq i \leq n$, ℓ_i is either constant or the identity function on $[t]$. An example of such a line in $[4]^5$ is the function

$$\ell(x) = (3, x, 1, x, 4)$$

whose image is the set of four points $\{(3, 1, 1, 1, 4), (3, 2, 1, 2, 4), (3, 3, 1, 3, 4), (3, 4, 1, 4, 4)\}$. A classic result in Ramsey theory, the Hales–Jewett Theorem [4], asserts that for all values of the parameters t and r, there exists a sufficiently large n such that any r-coloring of $[t]^n$ will contain a monochromatic combinatorial line (that is, a line such that the points in its image are all assigned the same color). The Hales–Jewett number $\mathrm{HJ}(t, r)$ is defined to be the least n which suffices.

A pigeonhole argument is enough to show that $\mathrm{HJ}(2, r) = r$ for all r. Moreover, Hindman and Tressler [5] have shown that $\mathrm{HJ}(3, 2) = 4$. For more difficult cases, no exact values are known, and the best upper bounds were shown by Shelah [6]. For $\mathrm{HJ}(4, 2)$, this upper bound is already enormous. Shelah proves that if $\mathrm{HJ}(t - 1, r) \leq n$, then $\mathrm{HJ}(t, r) \leq nf(n, r^{t^n})$, where $f(\ell, k)$ is an upper bound for a related problem that satisfies $k \uparrow\uparrow \ell \leq f(\ell, k) \leq k \uparrow\uparrow (2\ell)$. Starting from $\mathrm{HJ}(3, 2) = 4$, we get a bound for $\mathrm{HJ}(4, 2)$ between $2 \uparrow\uparrow 7$ and $2 \uparrow\uparrow 8$.

On the other hand, the best lower bounds for the Hales–Jewett numbers, which can be obtained from the van der Waerden theorem (as in [6]), are very far away from these upper bounds. For instance, the integers $\{1, 2, \ldots, 34\}$ can be 2-colored in such a way that no 4-term arithmetic progression is monochromatic [3]. This coloring can be used to define a coloring of $[4]^{11}$ with no monochromatic combinatorial line, by coloring a point $(x_1, x_2, \ldots, x_{11})$ with the color of $x_1 + x_2 + \cdots + x_{11} - 10$, which shows that $\mathrm{HJ}(4, 2) > 11$. In general, this argument yields merely exponential lower bounds on the Hales–Jewett numbers: Berlekamp [2] showed that for prime p, we can 2-color $p \cdot 2^p$ consecutive integers with no monochromatic p-term arithmetic progression. This opens up the possibility that the true values of the Hales–Jewett numbers may be much smaller.

We extend the boundary of which Hales–Jewett numbers are known to have reasonable values by proving the following result:

Theorem 1. *Whenever the* 10^{11}*-dimensional grid* $[4]^{10^{11}}$ *is 2-colored, there exists a monochromatic combinatorial line. That is,* $\mathrm{HJ}(4, 2) \leq 10^{11}$.

Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213
E-mail address: mlavrov@andrew.cmu.edu.

2. Setup

Given a combinatorial line $\ell : [4] \to [4]^n$, we define its *length* $|\ell|$ to be the number of coordinates $1 \le i \le n$ for which $\ell_i(x)$ varies with x. For example, the line given by $\ell(x) = (3, x, 1, x, 4)$ has length 2. (To justify this terminology, note that $|\ell|$ is the Hamming distance between the two endpoints $\ell(1)$ and $\ell(4)$, or indeed between any two points of ℓ.)

Fix a 2-coloring of $[4]^n$. For each length k, we classify the combinatorial lines of length k into three types, and count their densities:

- $p_2(k)$ is the fraction of lines of length k which have 2 points of each color.
- $p_3(k)$ is the fraction of lines of length k which have 3 points of one color, and 1 of the other.
- $p_4(k)$ is the fraction of monochromatic lines of length k.

We are also interested in the fraction of pairs of collinear points (on lines of length k) assigned the same color. On each line, there are 6 pairs of points. For lines counted by p_2, 2 pairs are monochromatic; for lines counted by p_3, 3 pairs are monochromatic; for lines counted by p_4, all 6 pairs are monochromatic. Therefore

$$q(k) := \frac{1}{3}p_2(k) + \frac{1}{2}p_3(k) + p_4(k) \tag{1}$$

counts the fraction of monochromatic pairs of points on lines of length k.

The grid $[4]^n$ contains large "cliques": sets of points in which any two are collinear. In such a clique, the number of monochromatic pairs is least when the colors are balanced, and even then is close to $\frac{1}{2}$. Thus, we expect that for at least some lengths k, $q(k) > \frac{1}{2} - \epsilon$ for some ϵ that goes to 0 with n. This intuition is correct in a way that we will make more precise.

If we could show the stronger statement that $q(k) > \frac{1}{2}$ for some k, the proof would be complete: $p_4(k)$ occurs in (1) with a coefficient greater than $\frac{1}{2}$, so we would know that $p_4(k) > 0$, which means that a monochromatic line exists.

Even if $q(k) < \frac{1}{2}$, a large value of $q(k)$ gives partial information: either $p_4(k) > 0$, or else $p_3(k)$ is close to 1. Therefore we can also prove that $p_4(k) > 0$ by showing that for some k, $q(k)$ is close to $\frac{1}{2}$, but $p_3(k)$ is bounded away from 1. More formally, we can solve (1) for $p_4(k)$ (substituting $p_2(k) = 1 - p_3(k) - p_4(k)$) to get

$$p_4(k) = \frac{3}{2}\left(q(k) - \frac{1}{6}p_3(k) - \frac{1}{3}\right). \tag{2}$$

We will prove that a monochromatic line exists by showing that the right-hand side of (2) is positive for some k.

3. Showing that $q(k)$ is close to $\frac{1}{2}$

It is hopeless to show that $q(k)$ approaches $\frac{1}{2}$ for any individual k. For example, the "checkerboard" coloring, which colors a point by the sum of its coordinates modulo 2, has $p_2(k) = 1$, and therefore $q(k) = \frac{1}{3}$, for all odd k. Instead, we prove an inequality for a weighted sum of the first κ values of $q(k)$, where κ is a parameter to be determined later.

3.1. A bound for n-dimensional hypercubes.

We begin by considering collinear pairs in the hypercube $[2]^n$. Here, lines consist of only 2 points, and therefore $q(k)$, defined as before, simply counts monochromatic lines of length k.

Lemma 3.1. *For every $\kappa \le n$, whenever $[2]^n$ is 2-colored,*

$$\sum_{k=1}^{\kappa}(\kappa - k + 1)q(k) \ge \frac{\kappa^2 - 1}{4}\left(1 - \kappa\sqrt{\frac{2}{\pi n}}\right). \tag{3}$$

Proof. The hypercube $[2]^n$ is the union of $n!$ chains of length $n+1$: maximal sequences of points $C_0, \ldots, C_n$ such that any two points C_i and C_j are collinear. For each permutation σ of $[n]$, we obtain such a chain by letting $C_i(\sigma)$ be the point with 2 in the coordinates $\sigma(1), \sigma(2), \ldots, \sigma(i)$ and 1 in all others. The line through $C_i(\sigma)$ and $C_j(\sigma)$, for $i < j$, has length $j-i$. Any permutation σ' such that $\{\sigma(1), \ldots, \sigma(i)\} = \{\sigma'(1), \ldots, \sigma'(i)\}$ and $\{\sigma(i+1), \ldots, \sigma(j)\} = \{\sigma'(i+1), \ldots, \sigma'(j)\}$ will satisfy $C_i(\sigma) = C_i(\sigma')$ and $C_j(\sigma) = C_j(\sigma')$; therefore $C_i(\sigma)$ and $C_j(\sigma)$ occur together in $i!(j-i)!(n-j)!$ chains.

Fix any 2-coloring of $[2]^n$. Let $Q_{i,j}(\sigma) = 1$ if $C_i(\sigma)$ and $C_j(\sigma)$ are given the same color, and 0 otherwise. Then the total number of monochromatic points on all $\binom{n}{k}2^{n-k}$ lines of length k is given by

$$\begin{aligned}\binom{n}{k}2^{n-k}q(k) &= \sum_{\sigma}\sum_{i=0}^{n-k}\frac{Q_{i,i+k}(\sigma)}{i!k!(n-i-k)!} \\ &= \frac{1}{n!}\binom{n}{k}\sum_{\sigma}\sum_{i=0}^{n-k}\binom{n-k}{i}Q_{i,i+k}(\sigma).\end{aligned}$$

Therefore

$$q(k) = \frac{1}{n!}\sum_{\sigma}\left(\sum_{i=0}^{n-k}\frac{\binom{n-k}{i}}{2^{n-k}}Q_{i,i+k}(\sigma)\right). \tag{4}$$

Let

$$q(k,\sigma) = \sum_{i=0}^{n-k}\frac{\binom{n-k}{i}}{2^{n-k}}Q_{i,i+k}(\sigma).$$

Then equation (4) shows that $q(k)$ is the average of $q(k,\sigma)$ over all σ. To complete the proof of Lemma 3.1, it suffices to show that for each permutation σ, inequality (3) holds with $q(k,\sigma)$ in place of $q(k)$.

Define $w_{i,j} := \frac{\binom{n-(j-i)}{i}}{2^{n-(j-i)}}$, a shorthand for the coefficient of $Q_{i,j}(\sigma)$ in $q(j-i,\sigma)$. We have

$$\sum_{k=1}^{\kappa}(\kappa-k+1)q(k,\sigma) \geq \sum_{h=0}^{n-\kappa}\left(\sum_{h\leq i<j\leq h+k} w_{i,j}Q_{i,j}(\sigma)\right) \tag{5}$$

because each term $w_{i,j}Q_{i,j}(\sigma)$ occurs in the right-hand side of (5) for at most $\kappa-(j-i)+1$ values of i; fewer if $j < \kappa$ or if $i > n-\kappa$.

Each sum

$$\sum_{h\leq i<j\leq h+k} w_{i,j}Q_{i,j}(\sigma)$$

counts (with varying weights) the number of monochromatic pairs among the $\kappa+1$ points $C_h(\sigma)$, $C_{h+1}(\sigma)$, …, $C_{h+\kappa}(\sigma)$, any two of which are collinear. There must be at least $2\binom{(\kappa+1)/2}{2} = \frac{\kappa^2-1}{4}$ such pairs; their number is minimized if half the points receive one color and half receive the other. We do not know which weights correspond to those pairs. However, at the very least, we have the lower bound

$$\sum_{h\leq i<j\leq h+k} w_{i,j}Q_{i,j}(\sigma) \geq \frac{\kappa^2-1}{4}w_h^*,$$

where w_h^* is the least of all weights $w_{i,j}$ for $h \leq i \leq j \leq h+\kappa$. (We allow $i = j$ to simplify calculations later, though such a weight does not occur in the sum.) Substituting this lower bound

into inequality (5), we get

$$\sum_{k=1}^{\kappa}(\kappa-k+1)q(k,\sigma) \geq \frac{\kappa^2-1}{4}\sum_{h=0}^{n-\kappa} w_h^*.$$

It remains to find a lower bound for the sum of the w_h^*.

From Pascal's identity it follows that $w_{i,j} = \frac{1}{2}(w_{i-1,j} + w_{i,j+1})$. Therefore each coefficient $w_{i,j}$ is a weighted average of some of the coefficients

$$w_{h,h}, w_{h,h+1}, \ldots, w_{h,h+\kappa}, w_{h+1,h+\kappa}, \ldots, w_{h+\kappa,h+\kappa}.$$

Since all of these coefficients are included in the minimum defining w_h^*, we know that w_h^* must be one of these. Furthermore, this sequence is unimodal, so $w_h^* = \min\{w_{h,h}, w_{h+\kappa,h+\kappa}\}$.

The sequence $w_{0,0}, w_{1,1}, \ldots, w_{n,n}$ is just the sequence $\frac{\binom{n}{0}}{2^n}, \frac{\binom{n}{1}}{2^n}, \ldots, \frac{\binom{n}{n}}{2^n}$, and is also unimodal. So the sum $\sum_{h=0}^{n-\kappa} w_h^*$ will begin by summing $w_{h,h}$ and eventually switch to summing $w_{h+\kappa,h+\kappa}$, skipping some κ terms. Therefore

$$\sum_{h=0}^{n-\kappa} w_h^* \geq \sum_{h=0}^{n} w_{h,h} - \kappa \max_{0\leq h\leq n} w_{h,h} = 1 - \kappa\frac{\binom{n}{\lfloor n/2\rfloor}}{2^n} \geq 1 - \kappa\sqrt{\frac{2}{\pi n}}.$$

It follows that

$$\sum_{k=1}^{\kappa}(\kappa-k+1)q(k,\sigma) \geq \frac{\kappa^2-1}{4}\left(1-\kappa\sqrt{\frac{2}{\pi n}}\right)$$

and by averaging this inequality over all permutations σ and applying equation (5), we obtain the desired inequality (3). □

3.2. **Extending the bound to the grid $[4]^n$.** By giving away another error term, we can extend Lemma 3.1 to all collinear pairs in $[4]^n$.

Lemma 3.2. *For every $\kappa \leq \frac{n}{4}$, whenever $[4]^n$ is 2-colored,*

$$\sum_{k=1}^{\kappa}(\kappa-k+1)q(k) \geq \frac{\kappa^2-1}{4}\left(1-e^{-(n-\kappa)/8}-3\kappa\sqrt{\frac{2}{\pi(n-\kappa)}}\right).$$

Proof. We say that a collinear pair of points $\ell(a), \ell(b)$ for some line ℓ and some $a, b \in [4]$ has *type* m if there are m coordinates in total in which either point is equal to a or b; in other words, $\ell_i(x)$ is the constant a or b for $m - |\ell|$ values of i. We define $q(k,m)$ to be the fraction of collinear pairs of type m and on lines of length k which are monochromatic.

The type of a collinear pair matters because a collinear pair of type m is contained in the m-dimensional subcube of $[4]^n$ obtained by letting all m coordinates of either point which are equal to a or b vary freely between the two values. In this m-dimensional subcube, two points are collinear if and only if the corresponding points of $\{a,b\}^m$ (obtained by dropping all coordinates not equal to a or b) are collinear, so it has the structure of the hypercube $[2]^m$. The fraction of collinear pairs in this subcube which are monochromatic satisfies Lemma 3.1. By averaging over all m-dimensional subcubes, which cover each collinear pair of type m exactly once, we obtain

$$\sum_{k=1}^{\kappa}(\kappa-k+1)q(k,m) \geq \frac{\kappa^2-1}{4}\left(1-\kappa\sqrt{\frac{2}{\pi m}}\right). \tag{6}$$

There are $6\binom{n}{k}4^{n-k}$ collinear pairs on lines of length k; of them, $\binom{m}{k}2^{m-k}$ are in each m-dimensional subcube, and there are $6\binom{n}{m}2^{n-m}$ such subcubes. So the fraction of lines of length k

which have type m is

$$\frac{\binom{m}{k}2^{m-k}\binom{n}{m}2^{n-m}}{\binom{n}{k}4^{n-k}} = \frac{\binom{n-k}{m-k}}{2^{n-k}}.$$

Therefore we may express $q(k)$ as a weighted average of all the $q(k,m)$ by

$$q(k) = \sum_{m=k}^{n} \frac{\binom{n-k}{m-k}}{2^{n-k}} q(k,m). \tag{7}$$

Unfortunately, the weight of $q(k,m)$ in this average depends on k as well as m, which prevents us from simply averaging inequality (6) over all m. To fix this problem, we replace the weights in (7) by lower bounds independent of k, which will result in an inequality relating $q(k)$ to $q(k,m)$. (We will assume that $1 \le k \le \kappa \le \frac{n}{4}$.)

For $m \le \frac{n}{4}$, our lower bound will be 0: we drop all terms where m is too low, because the statement of inequality (6) is too weak in such cases. Otherwise, we want to replace the weight by the minimum of $\binom{n-k}{m-k}2^{-(n-k)}$ over all $k \le \kappa$.

From Pascal's identity, we have $\binom{n}{r}2^{-n} = \frac{1}{2}\left(\binom{n-1}{r-1}2^{-(n-1)} + \binom{n-1}{r}2^{-(n-1)}\right)$. Applying this iteratively, we can express each $\binom{n-k}{m-k}2^{-(n-k)}$ as a weighted average of some of

$$\frac{\binom{n-\kappa}{m-\kappa}}{2^{n-\kappa}}, \frac{\binom{n-\kappa}{m-\kappa+1}}{2^{n-\kappa}}, \ldots, \frac{\binom{n-\kappa}{m}}{2^{n-\kappa}}.$$

This sequence is unimodal, so the minimum is achieved at one of the endpoints, and we may replace equation (7) by

$$q(k) \ge \sum_{m=n/4}^{n} \frac{\min\left\{\binom{n-\kappa}{m-\kappa}, \binom{n-\kappa}{m}\right\}}{2^{n-\kappa}} q(k,m). \tag{8}$$

Sum the inequality (6) over all $m \ge \frac{n}{4}$ with weights as in inequality (8). The right-hand side of (8) will be smallest when $m = \frac{n}{4}$, so we may use that value for all m. We obtain

$$\sum_{k=1}^{\kappa}(\kappa - k + 1)q(k) \ge \frac{\kappa^2 - 1}{4}\left(1 - \kappa\sqrt{\frac{2}{\pi n/4}}\right)\sum_{m=n/4}^{n} \frac{\min\left\{\binom{n-\kappa}{m-\kappa}, \binom{n-\kappa}{m}\right\}}{2^{n-\kappa}}. \tag{9}$$

It remains to simplify the right-hand side.

The omission of the first $n/4$ terms of the sum in (9) results in an error of $\sum_{m<n/4}\binom{n-\kappa}{m-\kappa}2^{-(n-\kappa)}$, which is simply the binomial probability $\Pr[\mathrm{Bin}(n-\kappa, \frac{1}{2}) < \frac{n}{4} - \kappa]$. By the Chernoff bound (see, e.g., [1]),

$$\Pr\left[\mathrm{Bin}(n-\kappa, \tfrac{1}{2}) < \frac{n}{4} - \kappa\right] < \Pr\left[\mathrm{Bin}(n-\kappa, \tfrac{1}{2}) < \frac{n-\kappa}{4}\right] \le \exp\left(-\frac{n-\kappa}{8}\right).$$

With these initial terms, the sum in (9) would be equal to 1, except for skipping κ terms near the middle, which occurs when the minimum switches from selecting $\binom{n-\kappa}{m-\kappa}$ to selecting $\binom{n-\kappa}{m}$. Each of these terms is at most $\binom{n-\kappa}{(n-\kappa)/2}2^{-(n-\kappa)} \le \sqrt{\frac{2}{\pi(n-\kappa)}}$, so we lose at most κ times this quantity. Therefore the sum in inequality (9) satisfies

$$\sum_{m=n/4}^{n} \frac{\min\left\{\binom{n-\kappa}{m-\kappa}, \binom{n-\kappa}{m}\right\}}{2^{n-\kappa}} \ge 1 - e^{-(n-\kappa)/8} - \kappa\sqrt{\frac{2}{\pi(n-\kappa)}}.$$

Combining the two error terms, we complete the proof. □

4. Showing that $p_3(k)$ cannot be arbitrarily close to 1

In this section, we say that a combinatorial line in a 2-colored grid $[4]^n$ is *odd* if it has an odd number of points of each color. That is, an odd line has 3 points of one color and 1 point of the other, so it is exactly the type of line counted by $p_3(k)$.

To bound $p_3(k)$ away from 1, we first find a set of lines in $[4]^4$ which cannot all be odd:

Lemma 4.1. *Whenever $[4]^4$ is 2-colored, the 15 lines*

$$\begin{array}{lll} \ell^1(x) = (x,2,3,4) & \ell^6(x) = (x,2,x,4) & \ell^{11}(x) = (x,x,x,4) \\ \ell^2(x) = (1,x,3,4) & \ell^7(x) = (x,2,3,x) & \ell^{12}(x) = (x,x,3,x) \\ \ell^3(x) = (1,2,x,4) & \ell^8(x) = (1,x,x,4) & \ell^{13}(x) = (x,2,x,x) \\ \ell^4(x) = (1,2,3,x) & \ell^9(x) = (1,x,3,x) & \ell^{14}(x) = (1,x,x,x) \\ \ell^5(x) = (x,x,3,4) & \ell^{10}(x) = (1,2,x,x) & \ell^{15}(x) = (x,x,x,x) \end{array}$$

cannot all be odd.

Proof. A key observation is that each point of $[4]^4$ lies on an even number of these lines. The point $(1,2,3,4)$ lies on the 4 lines of length 1, and no other. Take any other point (x_1,x_2,x_3,x_4) expressible as $\ell^j(x)$ for some index j and some $x \in \{1,2,3,4\}$. Any coordinate i where $x_i \neq i$ must be a variable coordinate of ℓ^j; any coordinate i where $x_i = i \neq x$ must be a constant coordinate of ℓ^j. There is always exactly one coordinate where $x_i = i = x$, so there are 2 choices for j, depending on whether that coordinate is variable or constant.

If $[4]^4$ is 2-colored, choose either of the colors, and add up the number of points of that color on each of the fifteen lines. This total must be even, because each point is counted an even number of times. However, 15 odd numbers cannot add up to an even total, so one of the lines must contribute an even number. Therefore not all 15 lines can be odd. □

Structures isomorphic to the set of lines $\ell^1,\ldots,\ell^{15}$ occur many times in $[4]^n$, and in each such structure at most $\frac{14}{15}$ of the lines are odd. So our next step is to show that by (more or less) averaging over all such structures, we get an upper bound of $\frac{14}{15}$ for the overall densities of odd lines of certain lengths, up to an error term.

Lemma 4.2. *For every $k \leq \frac{n}{4}$, whenever $[4]^n$ is 2-colored,*

$$\left(1 - \frac{16k^2}{n}\right)\left(\frac{4}{15}p_3(k) + \frac{6}{15}p_3(2k) + \frac{4}{15}p_3(3k) + \frac{1}{15}p_3(4k)\right) \leq \frac{14}{15}. \tag{10}$$

Proof. Fix a 2-coloring of $[4]^n$ and some $k \leq \frac{n}{4}$. Let a *k-embedding* of $[4]^4$ into $[4]^n$ be a function $L : [4]^4 \to [4]^n$ such that for each coordinate $1 \leq i \leq n$, L_i is either constant or given by $L_i(x_1,x_2,x_3,x_4) = x_j$ for some $j \in \{1,2,3,4\}$. Moreover, we require that for each j, there are exactly k coordinates in which L_i varies with x_j. Let $\mathcal{L}_k$ be the set of all k-embeddings $[4]^4 \to [4]^n$.

Each $L \in \mathcal{L}_k$ induces a 2-coloring of $[4]^4$, by taking the preimage under L of the coloring of $[4]^n$. Moreover, a line $\ell : [4] \to [4]^4$ corresponds to a line $L \circ \ell : [4] \to [4]^n$, with $|L \circ \ell| = k|\ell|$, which is odd if and only if ℓ is odd in the induced coloring.

Count the number of odd lines $L \circ \ell^j$, where $L \in \mathcal{L}_k$ and ℓ^j is one of the 15 lines of Lemma 4.1. For a fixed L, at most 14 of the lines $L \circ \ell^j$ are odd; therefore we count at most $14|\mathcal{L}_k|$ odd lines total.

Let $P_3(k)$ be the number of odd lines of length k in $[4]^n$, related to the density $p_3(k)$ by $P_3(k) = \binom{n}{k}4^{n-k}p_3(k)$. If each line of length k could be expressed $M(k)$ times as $L \circ \ell^j$, we would have the inequality

$$M(k)P_3(k) + M(2k)P_3(2k) + M(3k)P_3(3k) + M(4k)P_3(4k) \leq 14|\mathcal{L}_k|. \tag{11}$$

Unfortunately, the number of ways to express a line $\ell : [4] \to [4]^n$ as $L \circ \ell^j$ depends on ℓ; specifically, on the number of coordinates of ℓ with each constant value. Inequality (11) still holds, however, if we instead define $M(k)$ to be the minimum multiplicity of any line of length k.

We compute the minimum multiplicity for each of the four possible lengths. Below, let n_1, n_2, n_3, and n_4 denote the number of coordinates of ℓ with constant value 1, 2, 3, and 4, respectively.

- If $|\ell| = k$, then ℓ can be expressed as $L(x, 2, 3, 4)$ in $\binom{n_2}{k}\binom{n_3}{k}\binom{n_4}{k}$ ways; we also get the corresponding counts for expressions of the form $L(1, x, 3, 4)$, $L(1, 2, x, 4)$, and $L(1, 2, 3, x)$. In total the line is counted with multiplicity $\binom{n_2}{k}\binom{n_3}{k}\binom{n_4}{k} + \binom{n_1}{k}\binom{n_3}{k}\binom{n_4}{k} + \binom{n_1}{k}\binom{n_2}{k}\binom{n_4}{k} + \binom{n_1}{k}\binom{n_2}{k}\binom{n_3}{k}$. This sum is minimized when n_1, n_2, n_3, n_4 are as equal as possible, so
$$M(k) \geq 4\binom{\frac{n-k}{4}}{k}^3.$$
- If $|\ell| = 2k$, then ℓ can be expressed as $L(x, x, 3, 4)$ in $\binom{2k}{k}\binom{n_3}{k}\binom{n_4}{k}$ ways; we also get the corresponding counts for expressions of the form $L(x, 2, x, 4)$ and so on, for a total of $\binom{2k}{k}\left(\binom{n_3}{k}\binom{n_4}{k} + \binom{n_2}{k}\binom{n_4}{k} + \binom{n_1}{k}\binom{n_4}{k} + \binom{n_2}{k}\binom{n_3}{k} + \binom{n_1}{k}\binom{n_3}{k} + \binom{n_1}{k}\binom{n_2}{k}\right)$. This is, once again, minimized when n_1, n_2, n_3, n_4 are as equal as possible, so
$$M(2k) \geq 6\binom{2k}{k}\binom{\frac{n-2k}{4}}{k}^2.$$
- If $|\ell| = 3k$, then ℓ can be expressed as $L(1, x, x, x)$ or $L(x, 2, x, x)$ or $L(x, x, 3, x)$ or $L(x, x, x, 4)$ in a total of $\binom{3k}{k,k,k}\left(\binom{n_1}{k} + \binom{n_2}{k} + \binom{n_3}{k} + \binom{n_4}{k}\right)$ ways, so
$$M(3k) \geq 4\binom{3k}{k,k,k}\binom{\frac{n-3k}{4}}{k}.$$
- Finally, if $|\ell| = 4k$, then ℓ can only be expressed as $L(x, x, x, x)$, which can be done in
$$M(4k) = \binom{4k}{k,k,k,k}$$
ways.

We can further replace $|\mathcal{L}_k|$ by $\binom{n}{k,k,k,k,n-4k}4^{n-4k}$. This allows us to rewrite inequality (11) as

$$4\binom{\frac{n-k}{4}}{k}^3\binom{n}{k}4^{n-k}p_3(k) + 6\binom{2k}{k}\binom{\frac{n-2k}{4}}{k}^2\binom{n}{2k}4^{n-2k}p_3(2k) + \\ + 4\binom{3k}{k,k,k}\binom{\frac{n-3k}{4}}{k}\binom{n}{3k}4^{n-3k}p_3(3k) + \binom{4k}{k,k,k,k}\binom{n}{4k}4^{n-4k}p_4(4k) \leq \\ \leq \binom{n}{k,k,k,k,n-4k}4^{n-4k}.$$

This inequality can be simplified by factoring out $\frac{4^{n-4k}}{k!^4}$ from each term. If we also replace falling powers $r(r-1)(r-2)(\cdots)(r-s+1)$ by r^s on the right-hand side (as an upper bound) and by $(r-s)^s$ on the left-hand side (as a lower bound), we obtain

$$4(n-k)^k(n-5k)^{3k}p_3(k) + 6(n-2k)^{2k}(n-6k)^{2k}p_3(2k) + \\ + 4(n-3k)^{3k}(n-7k)^k p_3(3k) + (n-4k)^{4k}p_3(4k) \leq 14n^{4k}.$$

Finally, dividing through by n^{4k} yields factors such as $\left(1-\frac{k}{n}\right)^k$. By iteratively applying the inequality $(1-u)(1-v) \geq 1-u-v$ for $u, v \geq 0$, we bound the first such factor:

$$\left(1-\frac{k}{n}\right)^k \left(1-\frac{5k}{n}\right)^{3k} \geq \left(1-\frac{k^2}{n}\right)\left(1-\frac{15k^2}{n}\right) \geq 1-\frac{16k^2}{n}.$$

Similarly, the factors of $\left(1-\frac{2k}{n}\right)^{2k}\left(1-\frac{6k}{n}\right)^{2k}$, $\left(1-\frac{3k}{n}\right)^{3k}\left(1-\frac{7k}{n}\right)^{k}$, and $\left(1-\frac{4k}{n}\right)^{4k}$ are each at most $1-\frac{16k^2}{n}$. After pulling out this factor, we obtain the inequality (10). □

5. Completing the proof of Theorem 1

To simplify notation, let $p_3^+(k) := \frac{4}{15}p_3(k)+\frac{6}{15}p_3(2k)+\frac{4}{15}p_3(3k)+\frac{1}{15}p_3(4k)$ (the quantity bounded by Lemma 4.2) and let $q^+(k) := \frac{4}{15}q(k)+\frac{6}{15}q(2k)+\frac{4}{15}q(3k)+\frac{1}{15}q(4k)$. We noted previously that if $q(k)-\frac{1}{6}p_3(k) > \frac{1}{3}$ for some k, then equation (2) implies that $p_4(k) > 0$, so a monochromatic line exists. Similarly, showing that $q^+(k)-\frac{1}{6}p_3^+(k) > \frac{1}{3}$ is positive suffices: this is a weighted average, so $q(ik)-\frac{1}{6}p_3(ik) > \frac{1}{3}$ will hold for some $1 \leq i \leq 4$.

We express as much of the left-hand side of Lemma 3.2 as possible in terms of q^+. We assume that the still-undetermined parameter κ is a multiple of 4 for simplicity. In the sum

$$\sum_{k=1}^{\kappa/4}(\kappa+1-2k)q^+(k) \tag{12}$$

the coefficient of each $q(k)$ is 0 for $k > \kappa$, and otherwise maximized if k is divisible by both 3 and 4, in which case it is at most

$$\frac{4}{15}(\kappa+1-2\cdot k)+\frac{6}{15}\left(\kappa+1-2\cdot\frac{k}{2}\right)+\frac{4}{15}\left(\kappa+1-2\cdot\frac{k}{3}\right)+\frac{1}{15}\left(\kappa+1-2\cdot\frac{k}{4}\right) = \kappa+1-\frac{103}{90}k,$$

so it is always less than $\kappa+1-k$. This means pulling out the sum (12) from the left-hand side of Lemma 3.2 leaves each $q(k)$ with a positive coefficient: we may write

$$\sum_{k=1}^{\kappa}(\kappa+1-k)q(k) = \sum_{k=1}^{\kappa/4}(\kappa+1-2k)q^+(k)+\sum_{k=1}^{\kappa}R_k q(k) \tag{13}$$

where $R_1, \ldots, R_\kappa$ are all positive. Furthermore, though each R_k is tedious to calculate, since equation (13) is valid for all values of $q(1), \ldots, q(\kappa)$, it remains valid if we set each of them to 1, and therefore

$$\sum_{k=1}^{\kappa}R_k = \sum_{k=1}^{\kappa}(\kappa+1-k)-\sum_{k=1}^{\kappa/4}(\kappa+1-2k) = \frac{\kappa^2+\kappa}{2}-\frac{3\kappa^2}{16} = \frac{5\kappa^2+8\kappa}{16}.$$

If $q(k) > \frac{1}{2}$ for any k, then from equation (1) we can conclude that $p_4(k) > 0$ and a monochromatic line exists. So assume the contrary: that $q(k) \leq \frac{1}{2}$ for all k. Then equation (13) implies that

$$\sum_{k=1}^{\kappa}(\kappa+1-k)q(k) \leq \sum_{k=1}^{\kappa/4}(\kappa+1-2k)q^+(k)+\frac{5\kappa^2+8\kappa}{16}\cdot\frac{1}{2}.$$

Therefore, by applying Lemma 3.2,

$$\sum_{k=1}^{\kappa/4}(\kappa+1-2k)q^+(k) \geq \frac{\kappa^2-1}{4}(1-\epsilon(n,\kappa))-\frac{5\kappa^2+8\kappa}{32},$$

where $\epsilon(n,\kappa)$ is the relative error term

$$\epsilon(n,\kappa) := e^{-(n-\kappa)/8} + 3\kappa\sqrt{\frac{2}{\pi(n-\kappa)}}.$$

Dividing by $\frac{3\kappa^2}{16}$ to obtain a weighted average and simplifying, we are left with

$$\frac{16}{3\kappa^2}\sum_{k=1}^{\kappa/4}(\kappa+1-2k)q^+(k) \geq \frac{1}{2} - \frac{4}{3\kappa} - \frac{4}{3}\epsilon(n,k) + \frac{4(1-\epsilon(n,k))}{3\kappa^2}.$$

The last term is positive and may be dropped. Therefore there is some $k^* \leq \kappa/4$ for which $q^+(k^*) \geq \frac{1}{2} - \frac{4}{3\kappa} - \frac{4}{3}\epsilon(n,k)$.

On the other hand, Lemma 4.2 tells us that, as long as $4k^* < \sqrt{n}$, $p_3^+(k^*) \leq \frac{14}{15}\left(1 - \frac{16(k^*)^2}{n}\right)^{-1}$, which is at most $\frac{14}{15}\left(1 - \frac{\kappa^2}{n}\right)^{-1}$. Therefore a lower bound on $q^+(k^*) - \frac{1}{6}p_3^+(k^*) - \frac{1}{3}$ is

$$\frac{1}{6} - \frac{4}{3\kappa} - \frac{4}{3}\epsilon(n,k) - \frac{7}{45}\left(1 - \frac{\kappa^2}{n}\right)^{-1}, \tag{14}$$

which is valid for any $\kappa < \sqrt{n}$.

As $n \to \infty$ and $\kappa \to \infty$, provided that $\frac{\kappa^2}{n} \to 0$, (14) approaches $\frac{1}{90}$, and so a monochromatic line must exist. In particular, (14) is already positive for $n = 10^{11}$ and $\kappa = 368$, completing the proof. (More precisely, $n = 19\,012\,590\,257$ and $\kappa = 240$ are enough.)

References

[1] N. Alon and J. Spencer. *The Probabilistic Method*. John Wiley & Sons, Hoboken, NJ, 2008.

[2] E. R. Berlekamp. A construction for partitions which avoid long arithmetic progressions. *Canad. Math. Bull.*, 11:409–414, 1968.

[3] V. Chvátal. Some unknown van der Waerden numbers. In *Combinatorial Structures and their Applications (Proc. Calgary Internat. Conf., Calgary, Alta., 1969)*, pages 31–33. Gordon and Breach, New York, 1970.

[4] A. W. Hales and R. I. Jewett. Regularity and positional games. *Trans. Amer. Math. Soc.*, 106:222–229, 1963.

[5] Neil Hindman and Eric Tressler. The first nontrivial Hales-Jewett number is four. *Ars Combin.*, 113:385–390, 2014.

[6] Saharon Shelah. Primitive recursive bounds for van der Waerden numbers. *J. Amer. Math. Soc.*, 1(3):683–697, 1988.